Mixed forward-backward stability of the twolevel orthogonal Arnoldi method for quadratic problems

Karl Meerbergen
Javier Peréz

Report TW 680, July 2017

KU Leuven
Department of Computer Science
Celestijnenlaan 200A – B-3001 Heverlee (Belgium)

Mixed forward-backward stability of the twolevel orthogonal Arnoldi method for quadratic problems

Karl Meerbergen
Javier Peréz

Report TW 680, July 2017

Department of Computer Science, KU Leuven

Abstract

We revisit the numerical stability of the twolevel orthogonal Arnoldi (TOAR) method for computing an orthonormal basis of a second-order Krylov subspace associated with two given matrices. We show that the computed basis is close (on certain subspace metric sense) to a basis for a second-order Krylov subspace associated with nearby coefficient matrices, provided that the norms of the given matrices are not too large or too small. Thus, the results in this work provide for the first time conditions that guarantee the numerical stability of the TOAR method in computing orthonormal bases of second-order Krylov subspaces. We also study scaling the quadratic problem for improving the numerical stability of the TOAR procedure when the norms of the matrices are too large or too small. We show that in many cases the TOAR procedure applied to scaled matrices is numerically stable when the scaling introduced by Fan, Lin and Van Dooren is used.

Keywords : Krylov subspace, second-order Krylov subspace, Arnoldi algorithm, second-order Arnoldi algorithm, two-level orthogonal Arnoldi algorithm, numerical stability.
MSC : Primary : 65F15, 65F30,

MIXED FORWARD–BACKWARD STABILITY OF THE TWO–LEVEL ORTHOGONAL ARNOLDI METHOD FOR QUADRATIC PROBLEMS

KARL MEERBERGEN* AND JAVIER PÉREZ†

Abstract. We revisit the numerical stability of the two–level orthogonal Arnoldi (TOAR) method for computing an orthonormal basis of a second–order Krylov subspace associated with two given matrices. We show that the computed basis is close (on certain subspace metric sense) to a basis for a second–order Krylov subspace associated with nearby coefficient matrices, provided that the norms of the given matrices are not too large or too small. Thus, the results in this work provide for the first time conditions that guarantee the numerical stability of the TOAR method in computing orthonormal bases of second–order Krylov subspaces. We also study scaling the quadratic problem for improving the numerical stability of the TOAR procedure when the norms of the matrices are too large or too small. We show that in many cases the TOAR procedure applied to scaled matrices is numerically stable when the scaling introduced by Fan, Lin and Van Dooren is used.

Key words. Krylov subspace, second–order Krylov subspace, Arnoldi algorithm, second–order Arnoldi algorithm, two–level orthogonal Arnoldi algorithm, numerical stability

AMS subject classifications. 65F15, 65F30

1. Introduction. Given two complex matrices $A, B \in \mathbb{C}^{n\times n}$ and two starting vectors $r_{-1}, r_0 \in \mathbb{C}^n$, if we define the sequence $r_{-1}, r_0, r_1, \ldots, r_{k-1}$ by the recurrence relation

$$r_i = Ar_{i-1} + Br_{i-2}, \quad \text{for } i = 1, 2, \ldots, k-1,$$

then, the *second–order Krylov subspace* associated with A and B, introduced by Bai and Su [2], is the subspace

$$\mathcal{G}_k(A, B; r_{-1}, r_0) := \operatorname{span}\{r_{-1}, r_0, r_1, \ldots, r_{k-1}\}.$$

Projection methods based on second–order Krylov subspaces have been found to be reliable procedures for obtaining good approximations to the solutions of (structured) quadratic eigenvalue problems [2, 11], and for model order reduction of second–order dynamical systems [1, 11] and second–order time–delay systems [21]. These procedures start by computing an orthonormal set of vectors $\{q_1, q_2, \ldots, q_{k+1}\}$ such that

$$\operatorname{span}\{q_1, q_2, \ldots, q_{k+1}\} = \mathcal{G}_k(A, B; r_{-1}, r_0).$$

They continue by projecting the problem onto the subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$, reducing the size of the original problem. Finally, the projected problem is solved by using standard algorithms for small/medium–sized dense matrices. The convergence of these projection methods for quadratic eigenvalue problems is studied in [9].

The *second–order Arnoldi (SOAR)* method and the *two–level orthogonal Arnoldi (TOAR)* method [2, 11, 18], are two well–known algorithms for computing orthonormal bases of second–order Krylov subspaces. Both methods compute such bases

*Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium. Email karl.meerbergen@cs.kuleuven.be.

†Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium. Email: javier.perezalvaro@kuleuven.be. Supported by KU Leuven Research Council grant OT/14/074 and the Interuniversity Attraction Pole DYSCO, initiated by the Belgian State Science Policy Office.

by embedding the second–order Krylov subspaces into standard Krylov subspaces. Moreover, while the SOAR method is prone to numerical instability [11], the analysis performed in [11] provides solid theoretical evidence of the numerical stability of the TOAR method. More precisely, the TOAR method is backward stable in computing an orthonormal basis of the Krylov subspace in which the second–order Krylov subspace is embedded. In this work, we extend this result by showing that the computed orthonormal basis for the second–order Krylov subspace is close (in the standard subspace metric sense [16]) to a second–order Krylov subspace associated with matrices $A+\Delta A$ and $B+\Delta B$. This result is stated in Corollary 3.3, which is a consequence of the more general Theorem 3.1. These two results are the major contributions of this work. Additionally, in Section 4, we study how scaling the original quadratic problem affects the norms of ΔA and ΔB.

The notation used in the rest of the paper is as follows. We use lowercase letters for vectors and uppercase letters for matrices. In addition, we use boldface letters to indicate that a matrix (resp. vector) will be considered as a 2×1 block–matrix (resp. block–vector), whose blocks are denoted with superscripts as in

$$\mathbf{A} = \begin{bmatrix} A^{[1]} \\ A^{[2]} \end{bmatrix}.$$

The $n\times n$ identity matrix is denoted by I_n. By 0 we denote the zero matrix, whose size should be clear in the context. If $\mathcal{S},\mathcal{T}\subset\mathbb{C}^n$ are subspaces with the same dimension, say ℓ, we define the distance between $\mathcal{S}$ and $\mathcal{T}$ as

$$\operatorname{dist}(\mathcal{S},\mathcal{T}) := \|P_{\mathcal{S}} - P_{\mathcal{T}}\|_2, \tag{1.1}$$

where $P_{\mathcal{S}}$ and $P_{\mathcal{T}}$ are, respectively, orthogonal projectors onto $\mathcal{S}$ and $\mathcal{T}$. It is well–known that the distance function (1.1) is a metric on the set of all ℓ dimensional subspaces of $\mathbb{C}^n$ [16, Theorem 4.7]. Given a matrix $Q\in\mathbb{C}^{n\times k}$, with $k\le n$, we denote by $\operatorname{span}\{Q\}$ the subspace spanned by the columns of Q. We denote by $A^\dagger$ the Moore–Penrose pseudoinverse of a matrix A. By ϵ we denote the unit roundoff, and we use the notation $O(\epsilon)$ for any quantity that is upper bounded by ϵ times a modest constant.

2. The TOAR method and the stability analysis by Lu, Su and Bai. We review in this section the main ideas underlying the TOAR method for computing an orthonormal basis of a second–order Krylov subspace, and the result of the stability analysis performed by Lu, Su and Bai [11]. We essentially follow the presentation given in [11].

We begin by recalling that the second–order Krylov subspace $\mathcal{G}_k(A,B;r_{-1},r_0)$ can be embedded in a Krylov subspace associated with the companion matrix

$$C := \begin{bmatrix} A & B \\ I_n & 0 \end{bmatrix} \in \mathbb{C}^{2n\times 2n}. \tag{2.1}$$

Indeed, introducing the vector $\mathbf{v}_1 := \left[\begin{smallmatrix} r_0 \\ r_{-1}\end{smallmatrix}\right]$, the equality

$$\mathcal{K}_k(C;\mathbf{v}_1) := \operatorname{span}\left\{\mathbf{v}_1, C\mathbf{v}_1,\ldots,C^{k-1}\mathbf{v}_1\right\} = \operatorname{span}\left\{\begin{bmatrix} r_0 \\ r_{-1}\end{bmatrix}, \begin{bmatrix} r_1 \\ r_0\end{bmatrix},\ldots,\begin{bmatrix} r_{k-1} \\ r_{k-2}\end{bmatrix}\right\} \tag{2.2}$$

is immediately verified. Hence, if $\mathbf{V}_k\in\mathbb{C}^{2n\times k}$ is a matrix whose columns form a basis for $\mathcal{K}_k(C;\mathbf{v}_1)$ and writing

$$\mathbf{V}_k = \begin{bmatrix} V_k^{[1]} \\ V_k^{[2]}\end{bmatrix} \quad \text{with} \quad V_k^{[i]}\in\mathbb{C}^{n\times k}, \quad \text{for } i=1,2,$$

we readily obtain from (2.2) that

$$\operatorname{span}\left\{V_k^{[1]}\right\} = \operatorname{span}\{r_0, r_1, \ldots, r_{k-1}\} \quad \text{and} \tag{2.3}$$

$$\operatorname{span}\left\{V_k^{[2]}\right\} = \operatorname{span}\{r_{-1}, r_0, \ldots, r_{k-2}\}, \tag{2.4}$$

and, therefore,

$$\operatorname{span}\left\{\begin{bmatrix} V_k^{[1]} & V_k^{[2]} \end{bmatrix}\right\} = \mathcal{G}_k(A, B; r_{-1}, r_0).$$

Thus, introducing $d_k := \dim(\mathcal{G}_k(A, B; r_{-1}, r_0)) \leq k+1$ and denoting by $Q_k \in \mathbb{C}^{n \times d_k}$ a matrix whose columns form a basis for $\mathcal{G}_k(A, B; r_{-1}, r_0)$, we can write

$$\mathbf{V}_k = \begin{bmatrix} V_k^{[1]} \\ V_k^{[2]} \end{bmatrix} = \begin{bmatrix} Q_k U_k^{[1]} \\ Q_k U_k^{[2]} \end{bmatrix} =: (I_2 \otimes Q_k)\mathbf{U}_k, \tag{2.5}$$

for some matrix $\mathbf{U}_k \in \mathbb{C}^{2d_k \times k}$. We will refer to (2.5) as a *compact representation* of the matrix $\mathbf{V}_k$. Furthermore, from (2.5), we see that one (numerically expensive) possibility for computing a basis for the subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$ is by extracting it from a rank–revealing decomposition of the matrix

$$\begin{bmatrix} V_k^{[1]} & V_k^{[2]} \end{bmatrix}.$$

The TOAR method provides a stable and computationally–efficient alternative for computing such a basis.

Before introducing the TOAR method, let us recall that a matrix $\mathbf{V}_k$ whose columns form an orthonormal basis for $\mathcal{K}_k(C; \mathbf{v}_1)$ can be computed in a numerically stable way by applying the Arnoldi algorithm to the companion matrix (2.1). Certainly, in exact arithmetic, running k steps of the Arnoldi algorithm produces matrices satisfying

$$\begin{bmatrix} A & B \\ I_n & 0 \end{bmatrix} \mathbf{V}_k = \mathbf{V}_{k+1} \underline{H}_k, \tag{2.6}$$

where $\underline{H}_k \in \mathbb{C}^{(k+1)\times k}$ is an upper–Hessenberg matrix and the columns of $\mathbf{V}_k$ and $\mathbf{V}_{k+1} = \begin{bmatrix} \mathbf{V}_k & \mathbf{v}_{k+1} \end{bmatrix}$ form orthonormal bases for $\mathcal{K}_k(C; \mathbf{v}_1)$ and $\mathcal{K}_{k+1}(C; \mathbf{v}_1)$, respectively. We will refer to (2.6) as an *Arnoldi decomposition*.

By combining the compact representation (2.5) with the Arnoldi decomposition (2.6), we get the decomposition

$$\begin{bmatrix} A & B \\ I_n & 0 \end{bmatrix} (I_2 \otimes Q_k)\mathbf{U}_k = (I_2 \otimes Q_{k+1})\mathbf{U}_{k+1} \underline{H}_k, \tag{2.7}$$

which will be referred to as a *TOAR decomposition*. The TOAR method is a memory–efficient variant of the Arnoldi method [10, 12, 19] applied to the companion matrix (2.1) for computing (2.7). By exploiting the compact representation of $\mathbf{V}_k$ in (2.5)–(2.7), it computes matrices Q_k and $\mathbf{U}_k$ with orthonormal columns and the Hessenberg matrix $\underline{H}_k$, without forming explicitly the matrix $\mathbf{V}_k$. Notice in passing that the orthonormality of the columns of $\mathbf{U}_k$ and Q_k implies the orthonormality of the columns of $\mathbf{V}_k$. We refer the reader to [3, 11] for implementation details, and to [10, 13, 19] for extensions of the TOAR algorithm to other companion matrices.

In the presence of finite precision arithmetic, the TOAR method is numerically stable [11, 13], provided that the orthogonalization steps have been properly carried out [5]. More precisely, the computed matrices have, up to working precision, orthonormal columns and, together with the computed matrix $\underline{H}_k$, satisfy

$$R = \begin{bmatrix} A & B \\ I_n & 0 \end{bmatrix} (I_2 \otimes Q_k)\mathbf{U}_k - (I_2 \otimes Q_{k+1})\mathbf{U}_{k+1}\underline{H}_k, \tag{2.8}$$

for some matrix $R \in \mathbb{C}^{2n\times k}$ with $\|R\|_2 = O(\epsilon)\|C\|_2$. Then, it is standard to show from (2.8) that the columns of the computed matrix $\mathbf{V}_k = (I_2 \otimes Q_k)\mathbf{U}_k$ form a basis for a Krylov subspace $\mathcal{K}_k(C + E; \mathbf{v}_1)$, for some matrix E with $\|E\|_2 = O(\epsilon)\|C\|_2$ [11, 17]. However, the perturbation E destroys the companion matrix structure, i.e., the zero and identity blocks of C are not present in $C + E$. Therefore, it is not clear whether or not the columns of the computed matrix Q_k span a basis for some second–order Krylov subspace associated with perturbed matrices $A + \Delta A$ and $B + \Delta B$. This problem, left open in [11], is solved in the following section.

3. Mixed forward–backward stability of the TOAR method in computing an orthonormal basis of $\mathcal{G}_k(A, B; r_{-1}, r_0)$. The starting point is the residual (2.8), and our goal is to throw it back onto the matrices A and B. This is done in Theorem 3.1, which is one of our main results. As a corollary, we will obtain a mixed forward–backward stability result for the TOAR method in Corollary 3.3. The proof of Theorem 3.1 is postponed to the end of the section.

THEOREM 3.1. *Let $A, B \in \mathbb{C}^{n\times n}$ and let C be the companion matrix (2.1). Let $\underline{H}_k \in \mathbb{C}^{(k+1)\times k}$, and let $Q_k \in \mathbb{C}^{n\times d_k}$, $Q_{k+1} = \begin{bmatrix} Q_k & q_{k+1} \end{bmatrix} \in \mathbb{C}^{n\times(d_k+1)}$, with $d_k \le k + 1$, be full–column–rank matrices. Let*

$$\mathbf{U}_k = \begin{bmatrix} U_k^{[1]} \\ U_k^{[2]} \end{bmatrix} \in \mathbb{C}^{2d_k\times k} \quad \text{and} \quad \mathbf{U}_{k+1} = \begin{bmatrix} U_{k+1}^{[1]} \\ U_{k+1}^{[2]} \end{bmatrix} = \begin{bmatrix} U_k^{[1]} & x_k \\ 0 & \beta_k \\ U_k^{[2]} & y_k \\ 0 & 0 \end{bmatrix} \in \mathbb{C}^{2(d_k+1)\times(k+1)}$$

be also full–column–rank matrices. Let R be the residual (2.8), let $E = -R\mathbf{U}_k^{\dagger}(I_2 \otimes Q_k^{\dagger})$, and let $\mathcal{S}_k \subseteq \mathbb{C}^n$ be the subspace spanned by the columns of Q_k. If $\|E\|_2 < 1$, then there exists a d_k–dimensional second–order Krylov subspace $\mathcal{G}_k(A + \Delta A, B + \Delta B; \widetilde{r}_{-1}, \widetilde{r}_0)$ such that

$$\operatorname{dist}\left(\mathcal{S}_k, \mathcal{G}_k(A + \Delta A, B + \Delta B; \widetilde{r}_{-1}, \widetilde{r}_0)\right) \le \frac{\|E\|_2}{1 - \|E\|_2}, \tag{3.1}$$

for some vectors $\widetilde{r}_{-1}$ and $\widetilde{r}_0$, and some matrices ΔA and ΔB with

$$\|\Delta A\|_2 \le \|E\|_2 + \frac{\|E\|_2(1 + \|E\|_2)}{1 - \|E\|_2}, \tag{3.2}$$

and

$$\|\Delta B\|_2 \le \max\{1, \|A\|_2, \|B\|_2\}\left(\|E\|_2(2 + \|E\|_2) + \frac{\|E\|_2(1 + \|E\|_2)^2}{1 - \|E\|_2}\right). \tag{3.3}$$

REMARK 3.2. *The structure of the matrix $\mathbf{U}_{k+1}$ in Theorem 3.1 is imposed to make the matrix compatible with a TOAR decomposition [11, Lemma 31]. In*

particular, this is the structure of the computed matrix $\mathbf{U}_{k+1}$ *by the TOAR method in floating point arithmetic [11]. This structure for* $\mathbf{U}_{k+1}$ *will be assumed throughout the rest of the section.*

As an immediate corollary of Theorem 3.1, we obtain the following mixed forward–backward stability result for the TOAR method.

COROLLARY 3.3. *Let* $Q_k \in \mathbb{C}^{n\times d_k}$ *be the matrix obtained by the TOAR method run in a computer with unit roundoff equal to* ϵ*, i.e., the computed quantities satisfy* (2.8) *with* $\|R\|_2 = O(\epsilon)\|C\|_2$*. Assume that* Q_k *has full column rank, and let* $\mathcal{S}_k$ *be the subspace spanned by the columns of* Q_k*. Then, to first order in* ϵ*, there exists a* d_k*–dimensional second–order Krylov subspace* $\mathcal{G}_k(A+\Delta A, B+\Delta B; \widetilde{r}_{-1}, \widetilde{r}_0)$ *such that*

$$\operatorname{dist}\left(\mathcal{S}_k, \mathcal{G}_k(A+\Delta A, B+\Delta B; \widetilde{r}_{-1}, \widetilde{r}_0)\right) = O(\epsilon)\max\{1, \|A\|_2, \|B\|_2\}, \tag{3.4}$$

for some vectors $\widetilde{r}_{-1}$ *and* $\widetilde{r}_0$*, and some matrices* ΔA *and* ΔB *with*

$$\begin{aligned} &\|\Delta A\|_2 = O(\epsilon)\max\{1, \|A\|_2, \|B\|_2\}, \quad \textit{and} \\ &\|\Delta B\|_2 = O(\epsilon)\left(\max\{1, \|A\|_2, \|B\|_2\}\right)^2. \end{aligned} \tag{3.5}$$

In words, the columns of the computed matrix Q_k *span a subspace close to a second–order Krylov subspace associated with matrices* $A+\Delta A$ *and* $B+\Delta B$*.*

Proof. The error analysis by Lu, Su and Bai [11] shows that the columns of the matrices Q_k and $\mathbf{U}_k$ computed by the TOAR method are both well–conditioned bases of the subspaces they span, that is,

$$\|Q_k\|_2\|Q_k^\dagger\|_2 = 1 + O(\epsilon) \quad \text{and} \quad \|\mathbf{U}_k\|_2\|\mathbf{U}_k^\dagger\|_2 = 1 + O(\epsilon),$$

and that the norm of the residual (2.8) for the computed matrices is a modest multiple of the unit roundoff times the norm of the companion matrix. Setting $E = -R\mathbf{U}_k^\dagger(I_2 \otimes Q_k^\dagger)$, we obtain that Theorem 3.1 holds with

$$\|E\|_2 = O(\epsilon)\|C\|_2 = O(\epsilon)\max\{1, \|A\|_2, \|B\|_2\}. \tag{3.6}$$

To finish the proof, it suffices to notice that (3.1), (3.2) and (3.3), together with (3.6), imply (3.4) and (3.5) to first order in ϵ. ∎

Paraphrasing Higham [8], Theorem 3.1 tells us that, as long as the norms of the matrices A and B are not too large or too small, the basis computed by the TOAR method is "almost the right answer for almost the right data". Hence, in this situation, TOAR is numerically stable in computing orthonormal bases of second–order Krylov subspaces. When the norms of A and B are very large or very small, one could consider scaling the problem for improving the stability properties of TOAR. This is considered in Section 4.

The proof of Theorem 3.1 requires several technical results that we state in the following lemmas. Lemma 3.4 projects the residual (2.8) back on the companion matrix (2.1).

LEMMA 3.4. *If* R *denotes the residual* (2.8)*, then the matrix* $E = -R\mathbf{U}_k^\dagger(I_2 \otimes Q_k^\dagger)$ *satisfies the decomposition*

$$\left(\begin{bmatrix} A & B \\ I_n & 0 \end{bmatrix} + \underbrace{\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}}_{=E}\right)(I_2 \otimes Q_k)\mathbf{U}_k = (I_2 \otimes Q_{k+1})\mathbf{U}_{k+1}\underline{H}_k, \tag{3.7}$$

where E in (3.7) *has been partitioned conformably to the partition of the companion matrix* (2.1)*. In other words, the matrices satisfy an exact TOAR decomposition for a perturbed matrix $C + E$.*

Proof. It is immediately verified that $E(I_2 \otimes Q_k)\mathbf{U}_k = -R$. ∎

The decomposition (3.7) is an exact TOAR decomposition, but the matrix $C + E$ is not a companion matrix. Thus, we cannot yet associated Q_k with a second–order Krylov subspace. Nevertheless, Lemma 3.5 shows that, as long as the norm of the perturbation E in Lemma 3.4 is small enough, the companion structure of the perturbed companion matrix in (3.7) can be recovered via a similarity transformation.

LEMMA 3.5. *Let C be the companion matrix in* (2.1)*. If $E = \left[\begin{smallmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{smallmatrix}\right]$, where $E_{ij} \in \mathbb{C}^{n\times n}$, is a matrix such that $\|E_{21}\|_2 < 1$, then*

$$\begin{bmatrix} I_n & (I_n + E_{21})^{-1}E_{22} \\ 0 & (I_n + E_{21})^{-1} \end{bmatrix} (C + E) = \begin{bmatrix} A + \Delta A & B + \Delta B \\ I_n & 0 \end{bmatrix} \begin{bmatrix} I_n & (I_n + E_{21})^{-1}E_{22} \\ 0 & (I_n + E_{21})^{-1} \end{bmatrix},$$

where the matrices ΔA and ΔB are equal to

$$\Delta A = E_{11} + (I_n + E_{21})^{-1}E_{22}(I_n + E_{21}), \quad \textit{and} \tag{3.8}$$

$$\Delta B = BE_{21} + E_{12}(I_n + E_{21}) - (A + E_{11})(I_n + E_{21})^{-1}E_{22}(I_n + E_{21}). \tag{3.9}$$

In words, the companion structure can be recovered via a similarity transformation close to the identity.

Proof. The condition $\|E_{21}\|_2 < 1$ guarantees the nonsingularity of the matrix $I_n + E_{21}$. Then, the result can be easily checked by performing directly the matrix multiplications. ∎

REMARK 3.6. *The idea of recovering the "companion structure" of a perturbed companion matrix by using transformations close to the identity as in the proof of Lemma 3.5 has appeared several times in the context of studying the numerical stability of solving polynomial eigenvalue problems by linearization [14, 15, 20].*

Applying Lemma 3.5 to the perturbed companion matrix in the TOAR decomposition (3.7), we obtain

$$\begin{aligned} \begin{bmatrix} A + \Delta A & B + \Delta B \\ I_n & 0 \end{bmatrix} \begin{bmatrix} I_n & (I_n + E_{21})^{-1}E_{22} \\ 0 & (I_n + E_{21})^{-1} \end{bmatrix} (I_2 \otimes Q_k)\mathbf{U}_k \\ = \begin{bmatrix} I_n & (I_n + E_{21})^{-1}E_{22} \\ 0 & (I_n + E_{21})^{-1} \end{bmatrix} (I_2 \otimes Q_{k+1})\mathbf{U}_{k+1}\underline{H}_k, \end{aligned} \tag{3.10}$$

where the matrices ΔA and ΔB are defined in (3.8)–(3.9). Then, introducing the new matrices

$$\mathbf{W}_i = \begin{bmatrix} W_i^{[1]} \\ W_i^{[2]} \end{bmatrix} := \begin{bmatrix} Q_i U_i^{[1]} + (I_n + E_{21})^{-1}E_{22}Q_i U_i^{[2]} \\ (I_n + E_{21})^{-1}Q_i U_i^{[2]} \end{bmatrix}, \quad \text{for } i = k, k+1, \tag{3.11}$$

the decomposition (3.10) becomes

$$\begin{bmatrix} A + \Delta A & B + \Delta B \\ I_n & 0 \end{bmatrix} \begin{bmatrix} W_k^{[1]} \\ W_k^{[2]} \end{bmatrix} = \begin{bmatrix} W_{k+1}^{[1]} \\ W_{k+1}^{[2]} \end{bmatrix} \underline{H}_k. \tag{3.12}$$

The compact representation of the matrix $\mathbf{V}_k = (I_2 \otimes Q_k)\mathbf{U}_k$ is destroyed after premultiplying $\mathbf{V}_k$ by the matrix $\left[\begin{smallmatrix} I_n & (I_n+E_{21})^{-1}E_{22} \\ 0 & (I_n+E_{21})^{-1} \end{smallmatrix}\right]$. Nevertheless, the obtained decomposition (3.12) is an exact Arnoldi decomposition for a companion matrix. Hence,

we obtain

$$\operatorname{span}\left\{\begin{bmatrix} W_k^{[1]} & W_k^{[2]} \end{bmatrix}\right\} = \mathcal{G}_k(A+\Delta A, B+\Delta B; \widetilde{r}_{-1}, \widetilde{r}_0), \tag{3.13}$$

for some vectors $\widetilde{r}_{-1}, \widetilde{r}_0$ and some matrices $A+\Delta A$ and $B+\Delta B$. In Lemma 3.7, we obtain a basis for the subspace (3.13).

LEMMA 3.7. *Let* (3.7) *be the TOAR decomposition for a slightly perturbation of a companion matrix* C *as in* (2.1)*, and let* $\mathbf{W}_i$*, with* $i = k, k+1$*, be the matrices defined in* (3.11)*. If* $\|E_{21}\|_2 < 1$*, then*

$$\operatorname{span}\left\{\begin{bmatrix} W_k^{[1]} & W_k^{[2]} \end{bmatrix}\right\} = \operatorname{span}\left\{(I_n + E_{21})^{-1} Q_k\right\},$$

and $\dim\left(\operatorname{span}\left\{(I_n + E_{21})^{-1} Q_k\right\}\right) = d_k$.

Proof. From Lemma 3.5 and the hypothesis $\|E_{21}\|_2 < 1$, we obtain that the matrices $\mathbf{W}_i$, with $i = k, k+1$, satisfy the Arnoldi decomposition (3.12). Examining the bottom block of $\mathbf{W}_k$ in (3.11), we readily obtain that $\operatorname{span}\{W_k^{[2]}\} \subseteq \operatorname{span}\{(I_n + E_{21})^{-1} Q_k\}$. Further, from the bottom block of (3.12), together with (3.11), we obtain

$$W_k^{[1]} = (I_n + E_{21})^{-1} Q_{k+1} U_{k+1}^{[2]} \underline{H}_k.$$

Then, from the fact that the matrix $U_{k+1}^{[2]}$ is of the form (recall Remark 3.6)

$$U_{k+1}^{[2]} = \begin{bmatrix} U_k^{[2]} & y_k \\ 0 & 0 \end{bmatrix},$$

for some vector y_k, we obtain that the columns of $W_k^{[1]}$ are linear combinations of only the first d_k columns of $(I_n + E_{21})^{-1} Q_{k+1}$, i.e., the columns of $(I_n + E_{21})^{-1} Q_k$. Therefore, $\operatorname{span}\{W_k^{[1]}\} \subseteq \operatorname{span}\{(I_n + E_{21})^{-1} Q_k\}$. Finally, it is clear that $d_k = \operatorname{rank}(Q_k) = \operatorname{rank}((I_n + E_{21})^{-1} Q_k)$. ∎

The last auxiliary result for the proof of Theorem 3.1 is Lemma 3.8, which shows how a subspace spanned by the columns of a matrix A behaves under multiplicative perturbations of the matrix A.

LEMMA 3.8. [4, Theorem 3.3] *Let* $A \in \mathbb{C}^{m\times n}$ *and* $\widetilde{A} = (I_m + E)A \in \mathbb{C}^{m\times n}$*, where* $(I_m + E) \in \mathbb{C}^{m\times m}$ *is nonsingular. Then,*

$$\operatorname{dist}(\mathcal{S}, \widetilde{\mathcal{S}}) \leq \min\{\|E\|_2, \|(I_m + E)^{-1} E\|_2\},$$

where $\mathcal{S}$ *and* $\widetilde{\mathcal{S}}$ *are the subspaces spanned, respectively, by the columns of the matrices* A *and* $\widetilde{A}$*.*

We are finally in a position to prove Theorem 3.1.

Proof. **(of Theorem 3.1)**. Recall that $E = -R\mathbf{U}_k^{\dagger}(I_2 \otimes Q_k^{\dagger})$. Since $\|E\|_2 < 1$ and, thus, $\|E_{21}\|_2 < 1$, we obtain from Lemmas 3.4 and 3.5 that the matrices $\mathbf{W}_i$, with $i = k, k+1$, defined in (3.11) satisfy the Arnoldi decomposition (3.12). Therefore, (3.13) is a d_k–dimensional second–order Krylov subspace associated with matrices $A+\Delta A$ and $B+\Delta B$, with ΔA and ΔB as in (3.8) and (3.9), respectively. Furthermore, from (3.8), we obtain

$$\|\Delta A\|_2 \leq \|E\|_2 + \frac{\|E\|_2(1 + \|E\|_2)}{1 - \|E\|_2},$$

and from (3.9), we obtain

$$\|\Delta B\|_2 \leq \max\{1, \|A\|_2, \|B\|_2\} \left(\|E\|_2(2 + \|E\|_2) + \frac{\|E\|_2(1 + \|E\|_2)^2}{1 - \|E\|_2} \right),$$

where we have used $\|(I_n + E_{21})^{-1}\|_2 \leq (1 - \|E_{21}\|_2)^{-1}$ and $\|E_{ij}\|_2 \leq \|E\|_2$, for $i, j = 1, 2$, for obtaining both upper bounds.

From Lemma 3.7, we obtain that the columns of $\widetilde{Q}_k := (I_n + E_{21})^{-1} Q_k$ form a basis for the second–order Krylov subspace (3.13). Let $\widetilde{\mathcal{S}}_k$ be the subspace spanned by the columns of $\widetilde{Q}_k$. To finish the proof of Theorem 3.1, it suffices to bound the distance between the subspaces $\mathcal{S}_k$ and $\widetilde{\mathcal{S}}_k$ from above. Writing $Q_k = (I_n + E_{21})\widetilde{Q}_k$, we immediately obtain from Lemma 3.8

$$\operatorname{dist}(\mathcal{S}_k, \widetilde{\mathcal{S}}_k) \leq \min\{\|E_{21}\|_2, \|(I_n + E_{21})^{-1} E_{21}\|_2\} \leq \frac{\|E\|_2}{1 - \|E\|_2},$$

and the proof is completed. □

4. Scaling the quadratic problem for improving the numerical stability of the TOAR procedure. In this section, we study the effect of scaling the original quadratic problem on the stability of the TOAR procedure.

A reasonable definition of a stable algorithm for computing an orthonormal basis of a second–order Krylov subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$ is to require that the algorithm computes an exact basis of (or, up to machine precision, a basis close to) a second–order Krylov subspace associated with matrices $A + \Delta A$ and $B + \Delta B$, with ΔA and ΔB satisfying

$$\max\{\|\Delta A\|_2, \|\Delta B\|_2\} = O(\epsilon) \max\{\|A\|_2, \|B\|_2\},$$

or the more stringent condition

$$\max\left\{ \frac{\|\Delta A\|_2}{\|A\|_2}, \frac{\|\Delta B\|_2}{\|B\|_2} \right\} = O(\epsilon).$$

In the former case, we would say that the algorithm is normwise stable, and in the latter, coefficientwise stable.

From Corollary 3.3, we see that the TOAR method fails to be stable in two situations, namely, when the norms of A and B are much smaller or much bigger than 1. We show in this section that in any of these situations the computed subspace could gain in accuracy when using the TOAR method on an appropriate scaling of the quadratic problem.

Scaling the quadratic problem consists in replacing the matrices A and B by the matrices $A_\alpha := \alpha A$ and $B_\alpha := \alpha^2 B$, where α is a nonzero positive real number. This operation is reflected as a scaling of the eigenvalues of the quadratic eigenvalue problem or as a scaling of the frequencies of the transfer function of the second–order dynamical system. The companion matrix associated with A_α and B_α is

$$C_\alpha := \begin{bmatrix} A_\alpha & B_\alpha \\ I_n & 0 \end{bmatrix} = \alpha \begin{bmatrix} I_n & 0 \\ 0 & \alpha^{-1} I_n \end{bmatrix} \begin{bmatrix} A & B \\ I_n & 0 \end{bmatrix} \begin{bmatrix} I_n & 0 \\ 0 & \alpha I_n \end{bmatrix}. \tag{4.1}$$

According to the analysis in Section 3, applying the TOAR procedure to the companion matrix C_α produces a computed matrix Q_k satisfying a TOAR decomposition (recall Corollary 3.3) of the form

$$\begin{bmatrix} A_\alpha + \Delta A_\alpha & B_\alpha + \Delta B_\alpha \\ I_n & 0 \end{bmatrix} (I_2 \otimes \widetilde{Q}_k) \widetilde{\mathbf{U}}_k = (I_2 \otimes \widetilde{Q}_{k+1}) \widetilde{\mathbf{U}}_{k+1} \underline{H}_k, \tag{4.2}$$

for some matrices $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{U}}_{k+1}$, and some matrices ΔA_α and ΔB_α with

$$\begin{aligned}\|\Delta A_\alpha\|_2 &= O(\epsilon)\max\{1, \|A_\alpha\|_2, \|B_\alpha\|_2\}, \quad \text{and}\\ \|\Delta B_\alpha\|_2 &= O(\epsilon)\left(\max\{1, \|A_\alpha\|_2, \|B_\alpha\|_2\}\right)^2,\end{aligned}$$

and where $\widetilde{Q}_k$ is a matrix such that

$$\operatorname{dist}(\operatorname{span}(\widetilde{Q}_k), \operatorname{span}(Q_k)) = O(\epsilon)\max\{1, \|A_\alpha\|_2, \|B_\alpha\|_2\}.$$

Undoing the scaling by using (4.1), we obtain from (4.2) the perturbed TOAR decomposition

$$\begin{bmatrix} A + \alpha^{-1}\Delta A_\alpha & B + \alpha^{-2}\Delta B_\alpha \\ I_n & 0 \end{bmatrix} (I_2 \otimes \widetilde{Q}_k)\widehat{\mathbf{U}}_k = (I_2 \otimes \widetilde{Q}_{k+1})\widehat{\mathbf{U}}_{k+1}\underline{\widehat{H}}_k,$$

where one α has been absorbed by $\underline{H}_k$ and the other two α's have been absorbed by the bottom blocks of $\widetilde{\mathbf{U}}_k$ and $\widetilde{\mathbf{U}}_{k+1}$. We conclude that the computed matrix Q_k is such that the subspace spanned by its columns is within a distance

$$O(\epsilon)\max\{1, \alpha\|A\|_2, \alpha^2\|B\|_2\} \tag{4.3}$$

of a second–order Krylov subspace associated with matrices $A + \alpha^{-1}\Delta A_\alpha =: A + \Delta A$ and $B + \alpha^{-2}\Delta B_\alpha =: B + \Delta B$ with

$$\begin{aligned}\|\Delta A\|_2 &= O(\epsilon)\alpha^{-1}\max\{1, \alpha\|A\|_2, \alpha^2\|B\|_2\}, \quad \text{and}\\ \|\Delta B\|_2 &= O(\epsilon)\alpha^{-2}\left(\max\{1, \alpha\|A\|_2, \alpha^2\|B\|_2\}\right)^2.\end{aligned} \tag{4.4}$$

Hence, the problem of choosing an optimal scaling parameter α for improving the stability of the TOAR procedure is reduced to the problem of minimizing (4.3) and (4.4) over $\alpha \in \mathbb{R}^+$.

We attempt to find a good choice of the scaling parameter α by minimizing the function

$$f(\alpha) := \frac{1 + \|A\|_2\alpha + \|B\|_2\alpha^2}{\alpha},$$

which is essentially equivalent to minimizing (4.4). We find that $\alpha_{\text{opt}} := \|B\|_2^{-1/2}$ is a local minimum of $f(\alpha)$. This scaling parameter corresponds to the scaling introduced by Fan, Lin and Van Dooren [6] for improving the backward stability of solving quadratic matrix polynomials by linearization.

We summarize in Theorem 4.1 the effect of scaling the quadratic problem with $\alpha = \alpha_{\text{opt}}$ on the stability of computing an orthonormal basis of the second–order Krylov subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$ by applying the TOAR method to the scaled companion matrix $C_{\alpha_{\text{opt}}}$.

THEOREM 4.1. *Let $A, B \in \mathbb{C}^{n\times n}$, let $\alpha_{\text{opt}} = \|B\|_2^{-1/2}$, and let $Q_k \in \mathbb{C}^{n\times d_k}$ be the computed matrix by the TOAR method applied to the scaled companion matrix $C_{\alpha_{\text{opt}}}$ in a computer with unit roundoff equal to ϵ. Then, the following statements hold.*

(i) *If $\|A\|_2 \le \|B\|_2^{1/2}$, then the subspace spanned by the columns of Q_k is within a distance $O(\epsilon)$ of a second–order Krylov subspace associated with matrices $A + \Delta A$ and $B + \Delta B$ such that*

$$\|\Delta A\|_2 = O(\epsilon)\|B\|_2^{1/2} \quad \text{and} \quad \frac{\|\Delta B\|_2}{\|B\|_2} = O(\epsilon).$$

(ii) *If* $\|A\|_2 \approx \|B\|_2^{1/2}$*, then the subspace spanned by the columns of* Q_k *is within a distance* $O(\epsilon)$ *of a second–order Krylov subspace associated with matrices* $A + \Delta A$ *and* $B + \Delta B$ *such that*

$$\max\left\{\frac{\|\Delta A\|_2}{\|A\|_2}, \frac{\|\Delta B\|_2}{\|B\|_2}\right\} = O(\epsilon).$$

(iii) *If* $\|A\|_2 > \|B\|_2^{1/2}$*, then the subspace spanned by the columns of* Q_k *is within a distance* $O(\epsilon)\|B\|_2^{-1/2}\|A\|_2$ *of a second–order Krylov subspace associated with matrices* $A + \Delta A$ *and* $B + \Delta B$ *such that*

$$\frac{\|\Delta A\|_2}{\|A\|_2} = O(\epsilon) \quad \textit{and} \quad \|\Delta B\|_2 = O(\epsilon)\|A\|_2^2.$$

Proof. The results readily follow from (4.3) and (4.4) with $\alpha = \alpha_{\rm opt}$. ☐

From Theorem 4.1, we obtain that the TOAR procedure applied to the scaled companion matrix $C_{\alpha_{\rm opt}}$ is normwise stable in computing an orthonormal basis of $\mathcal{G}_k(A, B; r_{-1}, r_0)$ in the case where $\|A\|_2 \leq \|B\|_2^{1/2}$. When $\|A\|_2 \approx \|B\|_2^{1/2}$, the method is actually coefficientwise stable. However, when $\|A\|_2 \gg \|B\|_2^{1/2}$, the scaling with $\alpha = \alpha_{\rm opt}$ does not resolve the stability issues of the TOAR method. In the language of quadratic matrix polynomials, this situation corresponds to the so called heavily damped quadratic matrix polynomials [7], and it is still an open problem to devise simple scaling strategies for those.

Remark 4.2. *When* $\|A\|_2 \gg \|B\|_2^{1/2}$*, we could also consider the scaling with parameter* $\alpha = \|A\|_2^{-1}$*. In this case, we would obtain that the subspace spanned by the columns of the computed matrix* Q_k *by the TOAR method applied to* C_α *is withing a distance* $O(\epsilon)$ *of a second–order Krylov subspace associated with matrices* $A + \Delta A$ *and* $B + \Delta B$ *such that*

$$\frac{\|\Delta A\|_2}{\|A\|_2} = O(\epsilon) \quad \textit{and} \quad \|\Delta B\|_2 = O(\epsilon)\|A\|_2^2,$$

which is an improvement over part–(iii) *in Theorem 4.1 and Corollary 3.3.*

5. Conclusions. Second–order Krylov subspace projection methods combined with the TOAR procedure have demonstrated superior numerical results over the standard approaches based on linearization for the solution of quadratic eigenvalue problems and for model order reduction of second–order dynamical systems. In this work, we have shown that the computed basis by the TOAR method for the subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$ is, up to machine precision, a second–order Krylov subspace associated with nearby matrices $A + \Delta A$ and $B + \Delta B$, providing to the observed numerical superiority a solid theoretical foundation. We have also considered the effect of scaling the original quadratic problem on the numerical stability of the TOAR method in computing an orthonormal basis of $\mathcal{G}_k(A, B; r_{-1}, r_0)$, and showed that in many situations the TOAR procedure applied to a scaled companion matrix is normwise, or even coefficientwise stable, in computing such a basis.

REFERENCES

[1] Z. Bai, and Y. Su. Dimension reduction of large–scale second–order dynamical systems via second–order Arnoldi method. *SIAM J. Sci. Comput.*, 26, pp. 1692–1709 (2005).

[2] Z. Bai, and Y. Su. SOAR: A second–order Arnoldi method for the solution of the quadratic eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 26, pp. 640–659 (2005).
[3] C. Campos, and J. E. Roman. Parallel Krylov solvers for the polynomial eigenvalue problem in SLEPc. *SIAM J. Sci. Comput.*, 38(5), pp. S385–S411 (2016).
[4] N. Castro–Gonzalez, F. M. Dopico, and J. M. Molera. Multiplicative perturbation theory of the Moore–Penrose inverse and the least squares problem. Linear Algebra Appl., 503, pp. 1–25 (2016).
[5] J. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart. Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization. *Math. Comp.*, 30, pp. 772–795 (1976).
[6] H.–Y. Fan, W.–W. Lin, and P. Van Dooren. Normwise scaling of second order polynomial matrices. *SIAM J. Matrix Anal. Appl.*, 26(1), pp. 252–256 (2005).
[7] S. Hammarling, C. Munro, and F. Tisseur. An Algorithm for the complete solution of quadratic eigenvalue problems. *ACM Transactions on Mathematical Software*, 39(3), pp. 18:1–18:19 (2013).
[8] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. 1st edition, SIAM, Philadelphia, 1996.
[9] T.–M. Huang, Z. Jia, and W.–W. Lin. On the convergence of Ritz pairs and refined Ritz vectors for quadratic eigenvalue problems. *BIT Numer. Math.*, 53, pp. 941–958 (2013).
[10] K. Kressner, and J. E. Roman. Memory–efficient Arnoldi algorithms for linearizations of matrix polynomials in Chebyshev basis. *Numer. Linear Algebr.*, 21, pp. 569–588 (2014).
[11] D. Lu, Y. Su, and Z. Bai. Stability analysis of the two–level orthogonal Arnoldi procedure. *SIAM J. Matrix Anal. Appl.*, 37, pp. 195–214 (2016).
[12] K. Meerbergen. The quadratic Arnoldi method for the solution of the quadratic eigenvalue problem. *SIAM J. Matrix Anal. Appl.*, 30(4), pp. 1463–1482 (2008).
[13] K. Meerbergen, and J. Pérez. Error analysis of the two–level orthogonal Arnoldi method for solving linearized polynomial eigenvalue problems. In preparation (2017).
[14] Y. Nakatsukasa, and V. Noferini, *On the stability of computing polynomial roots via confederate linearizations*. *Math. Comp.*, 85 (301), pp. 2391–2425 (2016).
[15] V. Noferini, and J. Pérez. Chebyshev rootfinding via computing eigenvalues of colleague matrices: when is it stable? Math. Comp., 86, pp. 1741–1767 (2016).
[16] G. W. Stewart, and J. –G. Sun. *Matrix Perturbation Theory*, Academic Press, New York, 1990.
[17] G. W. Stewart. Backward error bounds for approximate Krylov subspaces. Technical report UMIACS TR–2001–32 CMSC TR–4247, University of Maryland, Institute for Advanced Computer Studies, Department of Computer Science (2001).
[18] Y. Su, J. Zhang, and Z. Bai. A compact Arnoldi algorithm for Polynomial Eigenvalue Problems. `http://math.cts.nthu.edu.tw/Mathematics/RANMEP%20Slides/Yangfeng%20Su.pdf`
[19] R. Van Beeumen, K. Meerbergen, and W. Michiels. Compact rational Krylov methods for nonlinear eigenvalue problems. *SIAM J. Matrix Anal. Appl.*, 36, pp. 820–838 (2015).
[20] P. Van Dooren, and P. Dewilde. The eigenstructure of an arbitrary polynomial matrix: computational aspects. *Linear Algebra Appl.*, 50, pp. 545–579 (1983).
[21] Y. Zhang, and Y. Su. A memory–efficient model order reduction for time–delay systems. *BIT Numer. Math.*, 53, pp. 1047–1073 (2013).